Research Signpost, 37/661 (2), Fort P.O., Trivandrum - 695 023, Kerala, India.

RESEARCH SIGNPOST

Recent Research Developments in Protein Folding, Stability and Design, 2002: ISBN: 81-7736-070-1
Editors: M. Michael Gromiha and S. Selvaraj

15 Critical Building Blocks in Proteins: A Common Theme for Folding and Function

Sandeep Kumar [1], Adi Barzilai [2], Nurit Haspel [2], Yuk Yin Sham [3], Chung-Jung Tsai [4], Haim J. Wolfson [5] and Ruth Nussinov [2,4]

[1] Laboratory of Experimental and Computational Biology, NCI-Frederick Bldg 469, Rm 151 Frederick, MD 21702, USA

[2] Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

[3] Biomolecular Simulation and Scability Group, Thomas J. Watson Research Center, IBM Corporation P.O. Box 218, Yorktown Heights, NY 10598, USA

[4] Intramural Research Support Program - SAIC Laboratory of Experimental and Computational Biology, NCI-Frederick Bldg 469, Rm 151, Frederick, MD 21702, USA

[5] School of Computer Science, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

Here we illustrate that folding and function may be related in some proteins. Recently, we have developed a building blocks model of protein folding. In this model, a proteins folds via binding of conformationally uctuating building blocks. In a protein structure, one (or more) building block(s) may be more important than others. In the absence of such critical building block(s), the protein may not be able to acquire its native state. We have developed an algorithm to identify critical building blocks in proteins. Our analysis indicates that critical building blocks are evolutionarily conserved and contain functional residues, suggesting that these segments of protein structure are important for both folding and function.

Correspondence/Reprint request: R. Nussinov at NCI-Frederick Bldg 469, rm 151, Frederick, MD 21702. USA TEL: 301-846-5579, FAX: 301-846-5598, E-mail: ruthn@ncifcrf.gov.

INTRODUCTION

To be functional, proteins must acquire unique structures. All the necessary information required for the protein to fold into its three dimensional structure is contained in its one dimensional amino acid sequence. However, how the sequence information species the structure is not yet completely understood. In the protein literature, this fundamental question is referred to as the 'protein folding problem'. For the beginner, it must be clari ed that folding is not a problem for the protein. Given the right conditions of solvent, pH, salt concentration, temperature, etc., most proteins fold spontaneously into their respective three dimensional structures within time scales of milliseconds to seconds. On the other hand, protein folding is a major unsolved problem for the scientists who are trying to understand its mechanism. Over the past few decades, several models have been proposed for protein folding [1]. However, no single model can fully explain the mechanism of protein folding.

An important aspect of the protein folding problem is to analyze how different segments along the amino acid sequence are arranged in the protein structure. It is conceivable that one (or a few) segment(s) on the polypeptide chain may be more important for the overall protein fold than others. What if the same segment which is important for protein folding is also important for function? This would indicate a coupling between protein folding and protein function. Such a conjecture is evolutionarily attractive, since it implies that nature needs to conserve the same protein region for both folding and function. We denote such segments critical building blocks. In this chapter, we summarize our initial studies on critical building blocks in proteins.

THE BUILDING BLOCKS MODEL FOR PROTEIN FOLDING

Recently, we have developed a building blocks model of protein folding. In this model, we describe a folded protein as consisting of a set of hydrophobic folding units. A hydrophobic folding unit is a compact substructure of the protein that buries a large enough hydrophobic core and is capable of an independent, thermodynamically stable existence [1-5]. Hydrophobic folding units may coincide with protein domains, or constitute their sub-parts. A hydrophobic folding unit is the outcome of combinatorial assembly of a set of building blocks. In our deffnition, a building block consists of 15 or more contiguous amino acids. Unlike a hydrophobic folding unit, an isolated building block may not be stable in solution. Hence, the conformation of a building block seen in the native protein structure may or may not be the one seen in solution in the absence of the other building blocks. Furthermore, for a given building block, no single conformation necessarily prevails in solution. We visualize protein folding as proceeding through the binding of these conformationally fluctuating building blocks with one another *via* population selection [1]. Recently, we have incorporated these ideas into a computer program [6]. The program depicts the anatomy of protein structures at various hierarchical levels. Using an iterative, top-down cutting process, a protein tertiary structure is cut to reveal domains, then hydrophobic folding units and finally a set of fluctuating building blocks. The resulting anatomy tree-like organization describes the most likely folding pathway, kinetics and susceptibility of the protein to misfold. This process also illustrates whether the polypeptide chain folds in a sequential or more complex manner [6]. The complexity of a protein fold can be described in terms of the arrangement of the building blocks in the protein tertiary structure. If building blocks adjacent in the amino acid sequence are also adjacent in the protein 3-D structure, folding can be classified as sequential. Otherwise, it is a non-sequential folding. Different levels of protein folding complexity can be described between these two folding types [7].

CRITICAL BUILDING BLOCKS IN PROTEINS

All building blocks and their combinatorial assemblies are required for the native protein fold. Nevertheless, formation of one (or a few) building block(s) may be more important for correct protein folding than formation of the others. This is particularly true for proteins that fold in complex non sequential manner [8]. In such proteins, a building block that is in contact with several other building blocks at different hierarchical levels of the protein anatomy tree may be critical for achieving the correct protein fold. To be critical a building block must fulfill three conditions. First, it should be in contact with most (or all) other building blocks. Second, it may be inserted between two sequentially connected (or neighboring) building blocks. And third, in the absence of this building block, the remaining protein acquires non-native conformation. In general, while mutations have little impact on the overall protein structure, some mutations have more drastic consequences [9-11]. However, since this region is important for correct protein folding, its amino acid sequence appears less tolerant to mutations. This is apparent from its higher sequence conservation.

Since we already had a computer program that can cut a given protein structure into a set of building blocks at different hierarchical levels, we have written an add-on program to examine if a protein contains a critical building block. The add-on program assigns a critical building block index (CIndex) to each building block, based upon its location in the protein globule, the identity and number of other building blocks it interacts with and the extent of its surface area buried by such interactions. The interactions are measured in terms of the polar and non-polar surface areas buried among the building blocks. These areas get additional weights if the building block in question mediates interactions among the other building blocks. The details of the procedure of identifying the critical building blocks are given in reference [8]. The significance of the CIndex values can be measured by their Z-scores. The Z-score measures the difference of the CIndex value of a building block at a given level of protein anatomy cutting from the average CIndex value for all the building blocks at that level. Hence, for a building block i at the j^{th} level of the protein anatomy cutting, its Z-score is given by

$$Z\text{-}score_{ij} = (CIndex_{ij} - AvCIndex_j) / \sigma_j \qquad (1)$$

where $AvCIndex_j$ and σ_j are the average and standard deviations of the CIndex values for the building blocks at the j^{th} level.

Figure 1 shows an example of the protein anatomy cutting for Hen Egg White lysozyme at different hierarchical levels. The critical building block at the levels 2 and 3 is shown in the red color.

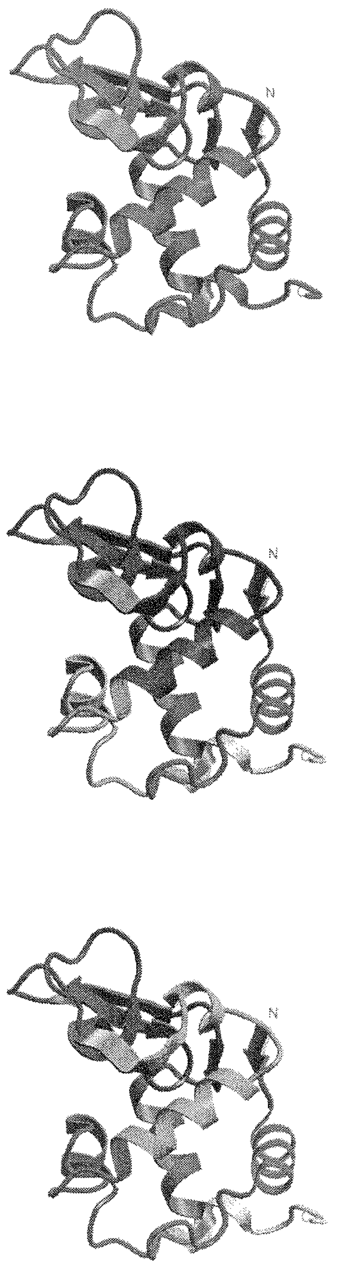


Figure 1: Diagram showing building block cutting of Hen Egg White Lysozyme (PDB entry: 135l) at levels 1 (top), 2 (middle) and 3 (bottom). In the middle and bottom panels, the building block (residues 2-39) shown in red color is the critical building block. This critical building block also contains Glu 35, an essential catalytic residue. Hence, this building block is critical for both function and folding of Hen egg white lysozyme.

We have attempted to identify the critical building blocks in a non-redundant data set of 938 non-homologous protein chains. 731 of these protein chains contain at least one building block with a CIndex value significant by $\geq 1\sigma$ at the lowest level of their anatomy cutting. 443 of these 731 protein chains have a critical building block with CIndex value significant by 1σ ($0.68 \leq p \leq 0.95$), 239 have a critical building block with a CIndex value significant by 2σ (($0.95 \leq p \leq 0.99$) and the remaining 49 proteins have a critical building block with a CIndex value significant by 3σ ($p>0.99$).

The CIndex values and their degrees of significance vary with protein size. Figure 2 shows the distribution of the critical building blocks at various levels of significance with respect to protein size. Smaller protein chains (up to 300 residues) usually have critical building blocks with CIndex values significant by 1σ. Mid-size protein chains (200 - 600 residues) contain critical building blocks with CIndex values significant by 2σ. Large protein chains (400 - 1000 residues) contain critical building blocks with CIndex values significant by 3σ. Hence, the use of an absolute significance level may not be a good idea to identify critical building blocks in proteins.

Figure 3 show the relationship between protein size, number of building blocks and the significance of the CIndex values for the critical building blocks. Figure 3(a) plots the number of building blocks at the lowest level of the protein anatomy cutting with respect to the protein size. As expected, the number of building blocks increases in bigger proteins. The maximum Z-score (the Z-score of the building block with the greatest CIndex value) at the lowest level of the protein anatomy cutting also increases with the protein size (Figure 3(b)). We have calculated another statistical measure, called t-score. The t-score for a building block at a given level of the protein anatomy cutting is obtained by multiplying its Z-score by the square root of the total number of

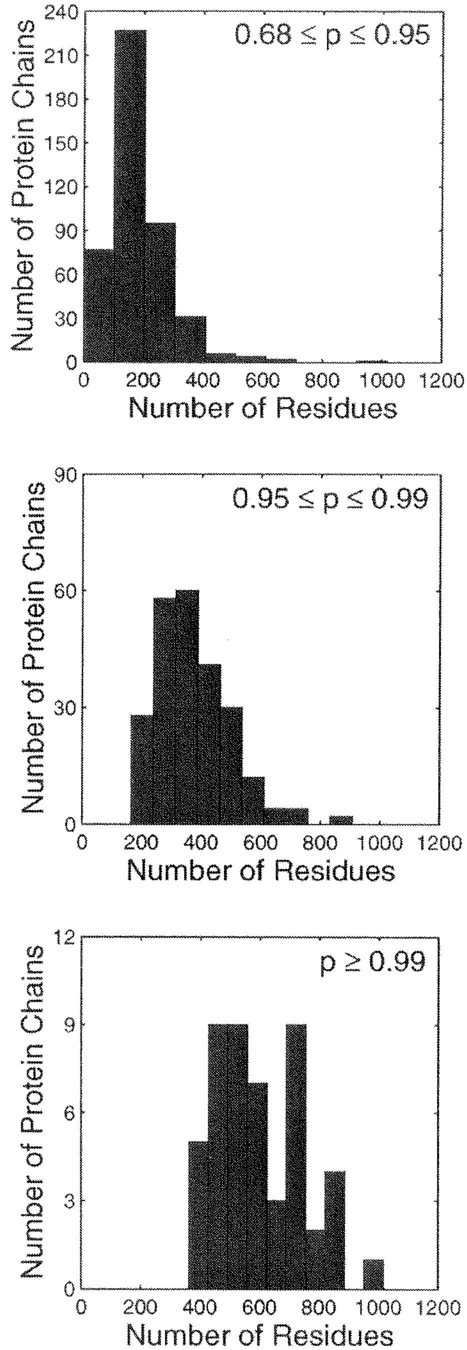


Figure 2: Histograms showing the distribution of protein chains containing critical building blocks with CIndex values significant by 1σ (top), 2σ (middle) and 3σ (bottom) with respect to size. The significance level of the critical building blocks depends upon the size of the protein chain.

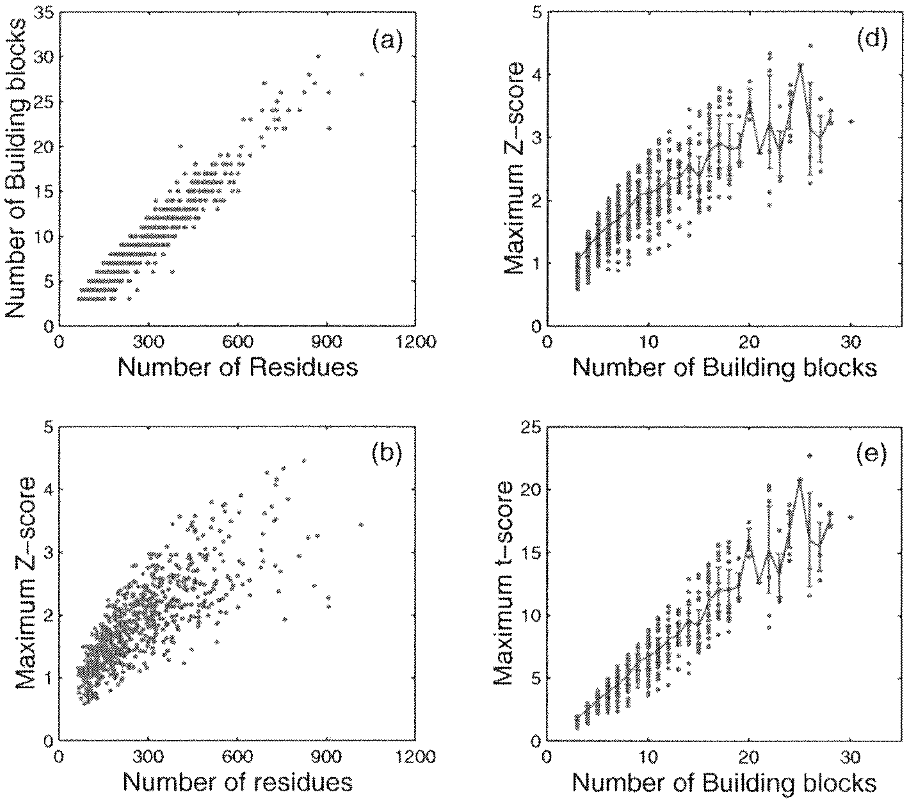
building blocks at that level. Figure 3(c) plots the maximum t-score at the lowest level of the protein anatomy cutting with the protein size. Again, the maximum t-score shows an increasing trend with the protein size. Since the criteria for identification of a critical building block are based on its interaction with other building blocks, we found the t-score a more useful test of significance for the building block CIndex values. 756 of 938 protein chains contain ≥1 building block with t-score greater than 1.0. At the lowest levels of the protein anatomy cutting, these protein chains contain a total of 6377 building blocks. We have used these proteins to construct a library of critical building blocks in proteins.

THE CRITICAL BUILDING BLOCKS LIBRARY

Figures 3(d) and 3(e) plot the Z-score and the t-score maxima at the lowest level of protein anatomy cutting as a function of the number of building blocks in the proteins. Consistent with the observations in Figures 3(a)-3(c), these increase with the number of building blocks. For each given number of building blocks, we have computed the average and standard deviation in the maximum Z-score and maximum t-score values. Using the maximum t-scores, we then identify the critical building block in a protein which contains *n* building blocks at the lowest level of the anatomy cutting as the one whose t-score is greater than the average plus 1σ of the maximum t-score expected for the *n* building blocks. Using these criteria, we are able to predict 225 protein chains as containing at least one critical building block. These constitute our critical building blocks library.

Many of the proteins in our critical building blocks library have been well studied and a large number of high resolution crystal structures are available for many of these proteins. The examples include Hen Egg White lysozyme (shown in Figure 1), avodoxin, sialidase, β-lactmase, tyrosine kinase, α-lytic protease, calmodulin, α-amylase, c-H Ras p21 protein, ferredoxin, diptheria toxin, rubisco, reverse transcriptase, lipase, class I MHC, porin, alkaline phosphatases, etc. Hence, it appears that a variety of proteins may contain critical building blocks.

Most of the proteins which are predicted to contain critical building blocks, fold in complex non-sequential manner. This is particularly true of mid-size and large proteins that have critical



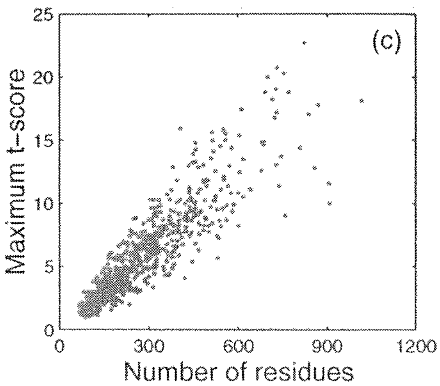


Figure 3: Plots showing the relationships among protein size, number of building blocks at the lowest level of the protein anatomy cutting and the significances of the CIndex values. These plots show (a) the number of building blocks *versus* the number of residues, (b) the maximum Z- score *versus* the number of residues, (c) the maximum t-score *versus* the number of residues, (d) the maximum Z-score *versus* the number of building blocks and (e) the maximum t-score *versus* the number of building blocks. The terms maximum Z-score and maximum t-score indicate the statistical significance of the greatest CIndex value assigned to a building block at the lowest level of the anatomy cutting by our procedure. In plots (d) and (e), the error bars (in green color) indicate the extents covered by 1σ about the average of the maximum Z-scores and maximum t-scores observed for *n* number of building blocks in proteins.

building blocks with CIndex values being significant by 2 or 3σ ($p>0.95$). Many of the critical building blocks lie at or near the N- or the C-termini of the polypeptide chains. Relatively large proteins may contain more than one building block with a significantly large CIndex value. In several such cases, different critical building blocks lie in different protein domains. In many cases, a single domain contains more than one critical building block. While all critical building blocks are located in the protein core, from our criteria defining a building block as critical it is evident that the presence of a building block in the core is an insufficient condition for a building block to be critical. Currently we are studying the proteins whose critical building blocks are in our library.

RELATIONSHIP BETWEEN PROTEIN FOLDING AND FUNCTION

To function, proteins must fold. This has been the central theme of protein science. Hence, most studies aim to interpret protein function in terms of speciftc interactions and environment of the protein structure. In recent years, investigators have reported examples of proteins existing in a 'disordered' state. Many proteins have been observed to be natively unfolded or contain disordered regions. Protein disorder has been thought to be related to protein function, especially for those involved in regulation and signal transduction [12-14 and references therein]. However, even in these examples, function requires the disordered regions to become ordered upon binding of their cognate partners. Although questioned by some [13], these observations actually uphold the validity of the structure-function paradigm for proteins. Consistent with this paradigm, functionally important residues are often found to be better conserved in evolution. However, other residues may also be strongly conserved. Multiple sequence aligments of c-type cytochromes and of proteins in globin families show that these non-functional conserved residues may act as folding nuclei for the proteins [15, 16]. Recently, Mirny and Shakhnovich have also demonstrated that residues which act as folding nuclei are significantly more conserved than other residues in the proteins [17]. Here we propose that residues which are important for function and residues which are important for correct protein folding fall in the same segment of the protein. Such segments are the critical building blocks.

In previous sections, we have focused on the important role of critical building blocks in folding. However, critical building blocks also contain functionally important residues. For example, the essential catalytic residue Glu-35 in Hen Egg lysozyme is located in the N-terminal critical building block shown in red color in Figure 1. Due to its central importance, the enzyme adenylate kinase has been extensively studied. Yeast adenylate kinase contains a critical building block at the N-terminus [8]. This critical building block is ~ 30 residues long and is conserved in adenylate kinases from other organisms such as *E. coli, B. stearothermphilus, Zea mays* and *Bos taurus* [8]. It also contains the phosphate binding loop (P-loop). The P-loop, also known as the giant anion hole, is characteristic of adenylate kinases and of a variety of ATP- and GTP- binding proteins [18-20]. The P-loop and the surrounding residues constituting the N-terminal building block show significant conservation in adenylate kinases [8]. Figure 4 shows the folding of yeast adenylate kinase. The critical building block is shown in red color. An N-terminal fragment (residues 1-36) in *E. Coli* dihydrofolate reductase (DHFR) is critical for its correct folding. This same fragment also forms an integral part of the active site of DHFR [21]. The fourth example is the pro-regions in α-lytic protease and in subtilisin. Pro-regions are critical for attaining the native fold. They also act as inhibitors for their corresponding proteases [22].

Currently, we are probing the critical building blocks in our library for their potential roles in protein function. The occurrence of residues that are important for both function and correct folding in the same region of the protein is an interesting observation. The residues in this region also show significant conservation. This makes sense from evolutionary point of view. For proteins that contain critical building blocks, it implies that guarding against mutations largely in a single segment may help protect both protein fold and function.

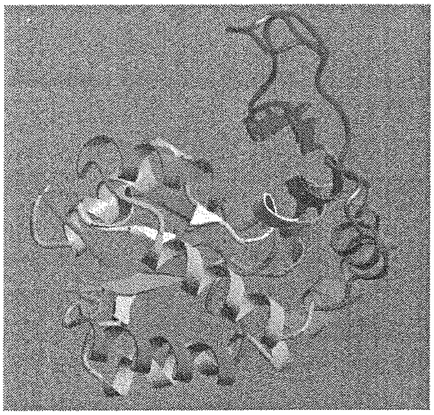


Figure 4: Diagram showing the building block cutting of yeast adenylate kinase (PDB entry: 1aky) at the lowest level of the protein anatomy. The critical building block (residues 3-32) is shown in red color.

PROTEIN FOLDING IN THE ABSENCE OF CRITICAL BUILDING BLOCKS

What happens to a protein, if its critical building block is removed? Removal of the critical building block from the protein core will expose the hydrophobic surface to water. One can imagine two potential scenarios. In first, the remainder of the protein may shrink to fill in the 'hole' created due to the removal of the critical building block. This may result in a stable non-native conformation for the remainder of the protein. In such a case, the conformations of the remaining building blocks will be unchanged, however, their associations will be non-native. In the second scenario, both the conformations of the other building blocks and their associations will be non-native. In extreme cases, the remainder of the protein may simply unfold. In any given case, the outcome will depend upon the extent of the hydrophobic surface exposed due to the removal of the critical building block.

We have performed molecular dynamics simulations of yeast adenylate kinase [8] with its first 36 residues (corresponding to the critical building block) removed. The simulations indicate that the remainder of the protein shrinks quickly to fill the 'hole' created by the absence of the critical building block. This results in a stable, more compact, non-native conformation. The other building blocks largely retain their native-like conformations. However, these building blocks mis-associate and consequently form non-native contacts. Since the P-loop is absent in the shrunk structure, the protein will be inactive. The removal of 'other' non-critical building blocks either does not perturb the native protein conformation or the extent of the perturbations is appreciably smaller [8].

THE STABILITY OF CRITICAL BUILDING BLOCKS

By definition, building blocks have fluctuating conformations in solution. Different building blocks may have different stabilities in solution. The stability of a building block may be estimated by an empirical scoring scheme [6]. In this scheme, each building block is assigned a score based on its compactness, hydrophobicity and isolatedness. In its isolated state, a critical building block would expose a large hydrophobic surface area that would have been otherwise buried due to contacts with other building blocks. The percentage of hydrophobic surface area exposed would be greater for a critical building block than for other building blocks. They are also less compact. Hence, critical building blocks are expected to be less stable as compared to other building blocks. A significant anti-correlation between CIndex values and building block scores is therefore expected.

Molecular dynamics simulations on the building blocks in *E. coli* DHFR at 300 K show that all are unstable. However, the N-terminal critical building block (residues 1-36) is highly unstable and unfolding is immediate for this building block when simulated alone. The other building blocks show more gradual unfolding [21].

SUMMARY AND OUTLOOK

Here we have proposed, and presented corroborating evidence that there may be a coupling between folding and function for some proteins. We have developed a building blocks model of protein folding. This model assumes hierarchical protein folding, and uses a hydrophobicity and compactness based criteria to identify the building blocks. At the end of several hierarchical cuttings of the native protein tertiary structure, we obtain a set of building blocks and a most likely protein folding pathway. For those proteins which fold in a sequential manner, all building blocks may be roughly equally important for the native structure. This may not be the case for proteins that fold in a complex, non-sequential manner. In such proteins, different building blocks may have different relative importance for the structure. One (or a few) building block(s) may be critical for the protein structure. In the absence of such a building block the protein may misfold. We have developed a simple computational procedure to identify such critical building blocks in proteins. Using a non-redundant database of non-homologous protein chains, we have developed a library of critical building blocks in 225 protein chains. Our present results are preliminary but encouraging. We have found the critical building blocks in Hen Egg White lysozyme, yeast adenylate kinase and *E. coli* dihydrofolate reductase to contain important active site residues. In the case of adenylate kinase, we observe that residues flanking the P-loop show a significantly greater sequence conservation. These residues and the P-loop form part of the N-terminal critical building block of adenylate kinase. This suggests that a protein which contains critical building blocks needs to guard against mutations in a single segment to protect both folding and function.

Critical building blocks may be disordered in solution in a manner similar to that observed for some regions in proteins. Several proteins, especially those involved in regulatory or signaling functions, have intrinsically disordered regions. Upon binding of their cognate DNA, metal ion, other protein molecule or small ligand/substrate, the disordered regions become ordered. Most often the disordered regions are also functionally important [12-14].

At present, our group is engaged in studying this structure-function coupling by using computational approaches. The observation of the coupling between protein folding and function *via* critical building blocks both furthers our understanding of protein folding, and may suggest ways for its utilization in protein folding schemes. It also raises several questions: Does the presence of critical building block(s) in a protein relate to its folding kinetics? Can we think of critical building blocks as potential folding nuclei?

Proteins containing critical building blocks can be expected to be more prone to mis-folding. Hence, identification of proteins containing such elements also *de facto* identifies proteins that are more likely to misfold.

With the availability of the critical building block library, we may be able to eventually identify critical building blocks in protein sequences. Application on a proteomic-scale, may help in identifying structurally and functionally important residues in proteins prior to their full structural and functional characterization.

ACKNOWLEDGMENTS

We thank Drs. Neeti Sinha, Kannan Gunasekaran, Buyong Ma, David Zanuy and, in particular, Jacob Maizel for numerous helpful discussions. Dr. Michael Gromiha is thanked for the invitation to write this article. The research of R. Nussinov and H. J. Wolfson in Israel has been supported in part by the Ministry of Science grant, and by the "Center of Excellence in Geometric Computing and its Applications" funded by the Israel Science Foundation (administered by the *Israel Academy of Sciences*). The research of H.J.W. is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily re ect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

REFERENCES

1. Tsai, C. J., Ma, B., Kumar, S., Wolfson, H. and Nussinov, R. 2001, Crit. Rev. Biochem. Mol. Biol., 36, 399-433.
2. Tsai, C. J., and Nussinov, R. 1997, Protein Sci., 6, 24-42.
3. Tsai, C. J., and Nussinov, R. 1997, Protein Sci., 6, 1246-1437.
4. Tsai, C. J., Xu, D. and Nussinov, R. 1998, Fold. & Des., 3, R71-R80.
5. Tsai, C. J., Kumar, S., Ma, B. and Nussinov, R. 1999, Protein Sci., 8, 1181-1190.
6. Tsai, C. J., Maizel Jr., J. V. and Nussinov, R. 2000, Proc. Natl. Acad. Sci. USA, 97, 12038-12043.
7. Tsai, C. J., Maizel, Jr., J. V., and Nussinov, R. 1999, Protein Sci., 8, 1591-1604.
8. Kumar, S. Sham Y. Y., Tsai, C. J. and Nussinov, R. 2001, Biophys. J., 80, 2439-2454.
9. Lim, W. A., and Sauer, R. T. 1991, J. Mol. Biol., 219, 359-376.
10. Lim, W. A., Farrugio, D. C., and Sauer, R. T. 1992, Biochemistry, 31, 4324-4333.
11. Matthews, B. W. 1993, Ann. Rev. Biochem., 62, 139-160.
12. Dunker, K. A. and Obradovic, Z. 2001, Nature Biotech., 19, 805-806.
13. Wright, P. E. and Dyson, H. J. 1999, J. Mol. Biol., 293, 321-331.
14. Tsai, C. J., Ma, B., Sham, Y. Y., Kumar, S. and Nussinov, R. 2001, Proteins: Struct. Funct. Genet., 44, 418-427.

15. Ptitsyn, O. B. and Ting, K. L. 1999, J. Mol. Biol., 291, 671-682.
16. Ptitsyn, O. B. 1998, J. Mol. Biol., 278, 655-666.
17. Mirny, L. and Shakhnovich, E. 2001, J. Mol. Biol., 308, 123-129.
18. Dreusicke, D. and Schulz, G. E. 1986, FEBS Lett. 208, 301-304.
19. Matte, A., Tari, L. W. and Delbaere, L. T. J. 1998, Structure, 6, 413-419.
20. Saraste, M., Sibbald, P. R., and Wittinghofer, A. 1990, Trends Biochem. Sci., 15, 430-434.
21. Sham Y. Y., Ma, B., Tsai, C. J., and Nussinov, R. 2001, Protein Sci., 10, 135-148.
22. Ma, B., Tsai, C. J. and Nussinov, R. 2000, Protein Eng., 13, 617-627.